

DOCUMENT RESUME

ED 093 328

IR 000 815

AUTHOR Jones, Martin V.; Jones, Richard M.
TITLE Lessons Learned in Past Social Service Program Evaluations, and Their Applicability to the NIE ESCD Demonstration.
INSTITUTION Impact Assessment Inst., Bethesda, Md.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C. Office of Programmatic Research and Development.
PUB DATE 15 May 74
NOTE 57p.; For related documents see IR 000 816-818
EDRS PRICE MF-\$0.75 HC-\$3.15 PLUS POSTAGE
DESCRIPTORS Case Studies; *Communication Satellites; Evaluation Criteria; *Evaluation Methods; Federal Programs; Program Evaluation; *Research Proposals; *Social Services
IDENTIFIERS Education Satellite Communications Demonstration; National Institute of Education; *NIE Design Proposal Review Board

ABSTRACT

The NIE (National Institute of Education) Design Proposal Review Board required advice to be used in selecting the Educational Satellite Communications Demonstration (ESCD) contractor. Therefore, a series of 25 "Lessons Learned and Problems Anticipated" (LLPA) sheets were prepared which seek to summarize the major insights relative to program evaluation methods that have been acquired from formal efforts to evaluate major social service programs in the last decade. Each of the sheets is subdivided into three sections plus a separate appendix sheet giving an evaluation problem, a particular social service program with this problem, a characteristic of the ESCD program relevant to this problem, and several authoritative quotations. The problems identified are generally methodological oversights, state-of-the-art limitations, generic program characteristics, and program administration. (WH)

ED 093328

BEST COPY AVAILABLE

LESSONS LEARNED IN PAST SOCIAL SERVICE PROGRAM EVALUATIONS,
AND THEIR APPLICABILITY TO THE NIE ESCD DEMONSTRATION

MARTIN V. JONES

RICHARD M. JONES

IMPACT ASSESSMENT INSTITUTE

May 15, 1974

IR 000 813

INTRODUCTION

The purpose of this report is to provide the NIE Design Proposal Review Board with some advice that it can use in selecting the Educational Satellite Communications Demonstration Phase III Assessment and Analysis contractor. The substance of the report is formatted in a series of 25 "Lessons Learned and Problems Anticipated" (LLPA) Sheets. These sheets seek to summarize the major insights relative to program evaluation methods that have been acquired from formal efforts to evaluate major social service programs in the last decade, both demonstration programs and operational programs.

Each of the 25 LLPA Sheets is subdivided into three sections plus a separate appendix sheet. The first section of each sheet provides a one paragraph generic statement describing one particular type of evaluation problem. The second section consists of a brief statement that references one or more particular social service program(s) in the past that encountered this problem in the course of its evaluation. The third section describes a particular characteristic of the ESCD program which could lead to this type problem developing in the course of the ESCD evaluation. Each appendix consists of one or

more quotations from an authoritative documented source that describes how the problem in question arose during the course of the project evaluation(s) cited in the second section.

Thinking in broad terms, the 25 types of evaluation problems identified on the LLPA Sheets can be classified into four categories:

1. Frequent Methodological Oversights. Many times because of ignorance, carelessness, or time constraints, an evaluation contractor will fail to employ the best professional practice.
2. State-of-the-Art Limitations. In many cases, no one--even the most skilled, meticulous professionals--knows how to handle very satisfactorily certain procedural problems in evaluating complex social service programs. A medical science analogue is the limited knowledge available for dealing with certain types of cancer and the common cold.
3. Generic Program Characteristics. Evaluation of a program can be difficult due to certain characteristics of the program itself. For instance, if the goals of a program, such as the Model Cities program, are highly diffuse, and inherently difficult to translate into measurable terms, the evaluation of that program becomes difficult.
4. Program Administration. In other cases, the way that a program is run once it starts, can complicate the evaluation task. For instance, if the program director decides to change the program's goals, ground rules, and activities frequently, the job of the evaluation task force is much more complicated than if a program is operated under more stable conditions.

The four-fold categorization described above is not discrete. Often a given type of evaluation problem might plausibly be classified under more than one of these categories. Where a problem is classified, however, is relatively unimportant.

The function of the classification scheme is primarily to serve as a communication device to help the Proposal Review Board, and other possible readers, to understand the basis on which our list of 25 types of evaluation problems was constructed.

There is no simple, straight-forward method for coping with any of the 25 problems identified in this report. For this reason, none of the three ESCD bidders can be expected to describe a completely adequate procedure for handling any of these problems. However, there are degrees of inadequacy. For instance, a bidder's technical plan may completely ignore a particular methodological problem, may treat it superficially, may discuss it in depth, or may provide a systematic plan for coping with the problem as well as discuss the characteristics of the problem, itself. Also, the bidder's management plan may disclose that the bidding organization has had no prior experience in dealing with a certain type of problem, little experience, moderate experience or

considerable experience. This management plan may also suggest that the bidder's proposed staff for the ESCD project has no, little, moderate, or considerable demonstrated capability for dealing with a particular type of problem.

The Proposal Review Board will probably want to consider other factors besides those covered in this report in rating the three bidders. For instance, one or more of the three bidders may be overcommitted on other projects that would prevent their doing justice to the ESCD project.

TABLE OF CONTENTS

<u>PROBLEM</u>	<u>PAGE</u>
1. Scoping the evaluation	1
2. Timing of evaluations	3
3. Ambiguous program goals	5
4. Distinguishing ends from means	7
5. Measures of program performance	9
6. Obtaining reliable data	11
7. Random selection of participants	13
8. Inadequate sampling	15
9. Open-ended interviews	17
10. Comparability: experimental and control groups	19
11. Knowledge transfers: experimental and control groups	21
12. Uncontrolled educational system characteristics	23
13. Quality of TV programming	25
14. Separating program vs. non-program effects	27
15. Detecting delayed, intended impacts	29
16. Anticipating unplanned side effects	31
17. Integrating incommensurables	33
18. Identifying Hawthorne effects	35
19. Duration of initial effects	37
20. Applying discount rates	39
21. Resistance to evaluation	41
22. Evaluator-target group interface	43
23. Midstream changes in program	45
24. Local diversity in demonstration programs	47
25. Replicability of findings	49
Expert Sources Contacted	51

Properly Scoping the Evaluation Effort

GENERAL STATEMENT OF THE PROBLEM

A difficult challenge in evaluating a major social service program--especially evaluations, such as the ESCD, that attempt to make a comprehensive "technology assessment" type analysis--is in properly "scoping" the effort. Most evaluations err in going to either of two extremes. Many evaluations are so narrow in their coverage that, at best, they are irrelevant because they give no consideration to certain important impacts. In contrast, other evaluations are overly ambitious in trying to cover every conceivable impact with the result that they end up treating everything superficially.

ILLUSTRATIVE PAST EXAMPLE

Headstart has been criticized for being concerned with only a narrow range of considerations, primarily cognitive and affective criteria. Conversely, the Belmont Project examined numerous problem areas, which, because of a lack of a clear focus, resulted in what some critics have termed "useless information."

ANTICIPATED ESCD ANALOGUE

The ESCD evaluation contractor, unless he carefully steers a middle ground, will very likely commit either of these historical errors. If his emphasis, like that in the Headstart evaluation, is on direct, immediate impacts, he will clearly miss the point of the whole undertaking. Conversely, if he tries to cover every impact area with equal emphasis, this evaluation, like others in the past, will prove nothing.

APPENDIX #1

Properly Scoping the Evaluation Effort

"One of the most frequent criticisms is the (Headstart) study is too narrow in scope, due to its central concern with cognitive and affective criteria. The Secretary of HEW at the time the report was released is reported to have immediately lodged this charge against the study."

Source: New York Times, April 25, 1969, p. 22.

"The fundamental problem (with the Belmont Project) has been that the system tries to accomplish too much. It asks too many questions about too many problem areas and, in doing so, produced almost totally useless information. Belmont would have been far more effective if there had been a clear decision at some early stage of its development to focus on a few key variables and to undertake accurately their measurement."

Source: Laurence E. Lynn, Jr. "Notes from HEW," Evaluation, Fall 1972, p. 28.

"It is important that the management and analytical staffs of an assessment project carefully consider the range of impact areas that will be analyzed and the depth to which this analysis will be carried. In an effort to be comprehensive, there is often a tendency to try to cover very many impact areas. However, because of time, money, and available data constraints, this effort to be comprehensive may sometimes be self-defeating--leading to a cursory, highly subjective coverage of many areas."

Source: MITRE Corporation M73-62 (M.V. Jones), "A Comparative State-of-the-Art Review of Selected U. S. Technology Assessment Studies," May 1973, p. 74.

The Timing of Evaluations

GENERAL STATEMENT OF THE PROBLEM

Often the time constraints under which an evaluation is conducted create methodological problems that would not otherwise exist. For instance, the early days of a new social service program are often characterized by much confusion. Hence, if an evaluation commences concurrently with the start of the demonstration program, itself, two things must be done. First, the evaluation procedure must be adjusted to discount the relevance of initial program effects that may disappear after the demonstration program administratively settles down. Second, the lessons learned should be clearly identified and documented so that program administrators and program evaluators can subsequently profit from this initial experience.

ILLUSTRATIVE PAST EXAMPLE

Some observers believe that the negative evaluation of the Head Start program was partly due to the evaluation commencing too soon, thus ignoring factors such as: the program's initial instability, inadequate physical facilities, and incomplete filling of professional positions.

ANTICIPATED ESCD ANALOGUE

Evaluation information gathered in the early stages of the ESCD program may be unrepresentative of the program's ultimate impacts because the evaluation will begin while the demonstration program, itself, is still in its formative stage.

The Timing of Evaluations

"The question of program potential leads to the issue of the timing of an outcome evaluation. In the case of Head Start, it may be argued that the program is still too new for its potential usefulness to be judged; and that it should be evaluated only when its operations have been stabilized, and properly trained professional staff and fully-equipped plants made available. Adequately trained personnel are not at present available in sufficient numbers to staff Head Start classrooms; in some regions they are not available at all. The very concept of Head Start, by design, called for experimentation with professional roles, new ways of teaching children whose backgrounds and skills were strange to most preschool teachers, much more intensive involvement with parents and the community, and new methods of plant utilization and recruiting children. All these considerations indicate that the program is bound, initially, to go through a formative, transitional period, thus seriously calling into question the timing of the evaluation to which this complicated program is being subjected so early in its evolution."

Source: Herbert Zimiles, "An Analysis of Current Issues in the Evaluation of Educational Programs," in Disadvantaged Child, Head Start and Early Intervention, (Volume 2), Jerome Hellmuth, Editor, Brunner/Mazel Publishers, N. Y., p. 548.

Ambiguous Program Goals

GENERAL STATEMENT OF THE PROBLEM

It is very difficult for even a well designed evaluation to determine whether the program was successful if a program's goals are stated in highly generalized terms.

ILLUSTRATIVE PAST EXAMPLE

Evaluations of broad aim programs, in which the objectives are not clearly stated, are difficult. An example is the Model Cities Program where the diverse objectives of the program were not explicitly defined.

ANTICIPATED ESCD ANALOGUE

The ESCD demonstration program has multiple goals, some of the most important ones being highly diffuse. The program's major objective will not be to ascertain whether satellite telecommunications can bring about improvements in the cognitive skills of students or in specific teaching skills of teachers. Rather it will be to ascertain in very broad terms whether satellites are a "feasible" medium for distributing educational materials to isolated rural populations. The difficulty is compounded by the fact that each of the three demonstration components has its own rather elusive set of goals.

APPENDIX #3

Ambiguous Program Goals

"A social welfare program (or for that matter any program) which does not have clearly defined goals cannot be evaluated without specifying some measurable goals.

. . . the desired end results of the Model Cities Program cannot be specified very clearly without radical simplification, a process which not only poses almost insurmountable obstacles to evaluators but leaves open the question whether any evaluation that is undertaken is an appropriate one."

Source: Peter H. Rossi and Walter Williams, Evaluating Social Programs: Theory, Practice, and Politics, Seminar Press, 1972, pp. 18-19.

LESSONS LEARNED AND PROBLEMS ANTICIPATED SHEET #4

Distinguishing Ends from Means

GENERAL STATEMENT OF THE PROBLEM

A major source of confusion that makes the evaluation of social-service programs difficult are the disputes among program sponsors, program administrators, program evaluators, and target group beneficiaries as to whether a particular type of activity represents an ultimate goal or end of the program or merely a means to the accomplishment of a larger end.

ILLUSTRATIVE PAST EXAMPLE

Community Action Programs were particularly subject to this source of confusion, especially when it concerned the matter of citizen participation. Administrators and evaluators were in disagreement, (1) whether citizen participation was a means or an end in itself and (2) how to define citizen participation, as it manifested itself in a variety of different ways.

ANTICIPATED ESCD ANALOGUE

Many facets of the ESCD experiments are likely to be subject to this "ends-means" debate. For instance, the secondary evening audiences in the Rocky Mountain states will be "interested professionals, parents, and other members of the community." It is very likely to be debated whether mere audience size is a sufficient criterion of success, with the question of whether the programming viewed affected the lives of the viewers in any significant way left unanswered. Similarly, the evaluation of career education training programming directed at teachers in both the Rocky Mountain and Appalachia regions is also likely to be caught up in a methodological debate. The ESCD evaluation contractor will have the difficult choice of deciding whether teacher exposure to the programming, itself, is a sufficient goal or should this be viewed merely as a means to determining whether there is any resultant impact on student choices relative to career goals.

APPENDIX #4

Distinguishing Ends from Means

"... in the Community Action Program it is not at all clear whether such things as resident participation or community organization are to be viewed as goals or as means. If the former, what is to be scored as participation: walking through the door of a local Community Action Agency, attending a meeting, serving on a committee, being in a policy-making position? Is it possible to establish a base line so as to assess relatively good or bad levels of participation? If participation is a means, what is the goal--a happier person, a more productive one?"

Source: Walter Williams, Social Policy Research and Analysis: The Experience in the Federal Social Agencies, American Elsevier Publishing Co., New York, N. Y., 1971, p. 86.

LESSONS LEARNED AND PROBLEMS ANTICIPATED SHEET #5

Developing Measures of Program Performance

GENERAL STATEMENT OF THE PROBLEM

Most formal evaluations of educational programs have been criticized because they did not use better measures of program performance. Usually the criticism has been more of a commentary on the methodological state-of-the-art and on the validity of drawing definite conclusions on the basis of questionable measurements, rather than a criticism that the evaluator failed to use the best professional practice that was available.

ILLUSTRATIVE PAST EXAMPLE

The Westinghouse Head Start Evaluation was criticized for using testing instruments which were not suitable to be applied to disadvantaged children, and which were probably constructed so as to be insensitive to changes that might occur in the children's cognitive and affective states.

ANTICIPATED ESCD ANALOGUE

In a larger sense the ESCD evaluation is directed more toward assessing the overall potentialities of satellite educational television than at improvements in specific cognitive skills or affective states. However, inevitably much of the evidence for the conclusions drawn from the observations made relative to improvements in cognitive skills (e.g., the skills of Appalachia teachers for teaching reading) or in affective states (e.g., attitudes, temperaments, and personality relative to career education of junior high school students in the Rocky Mountain states). Unless the Phase III evaluation contractor can surmount these measurement problems that evaluations of past educational programs have encountered, the ESCD evaluation will also be criticized for the inadequacy of its performance measures.

APPENDIX #5

Developing Measures of Program Performance

"The test instruments used in this study, and indeed all existing instruments for measuring cognitive and affective states in children, are primitive. They were not developed for disadvantaged populations, and they are probably so gross and insensitive that they are unable to pick up many of the real and important changes that Head Start has produced in children.

It is entirely possible this is true. However, most of the cognitive measures are the same ones being used by other child-development and Head Start researchers doing work on disadvantaged children. In those cases (relatively few) where previous studies have shown positive changes on these very same measures, they have seldom been questioned or disregarded because of the inadequacy of the instruments. In the affective area, Westinghouse found no appropriate test instruments and had to devise its own. Hence, the results should be viewed as suggestive, but no more."

Source: Walter Williams and John W. Evans, "The Politics of Evaluation: The Case of Head Start," Annals of the American Academy of Political and Social Science, September 1969, pp. 118-32.

LESSONS LEARNED AND PROBLEMS ANTICIPATED SHEET #6

Obtaining Complete, Reliable Data: Technical Problems

GENERAL STATEMENT OF THE PROBLEM

Credible program evaluation in the past has often been seriously hampered because the evaluation task force found it impossible to obtain reliable estimates of even the most basic socio-economic considerations.

ILLUSTRATIVE PAST EXAMPLE

The evaluations of HEW programs such as Vocational Rehabilitation, Adult Basic Education and Vocational Education encountered data problems in inadequate information about such matters as the cost per participant, data about age and sex distributions, etc. The Follow Through Evaluation suffered from a lack of reliable data owing to the method of field supervision.

ANTICIPATED ESCD ANALOGUE

Data limitations have already developed in the Alaskan ESCD component since apparently certain potential experimental sites were rejected as they would not provide a sufficient data sample. Even more important, the communities in all of these areas, and the people living in them, are much less experienced in conducting and participating in statistical surveys than are those in major urban centers. Perhaps, in the ESCD evaluation it will be necessary to place greater reliance on already existing records, rather than on special surveys, to obtain needed socio-economic data. However, existing records may also provide relatively unreliable baseline data concerning the socio-economic characteristics of target group participants.

APPENDIX #6

Obtaining Complete, Reliable Data: Technical Problems

"The process of analyzing and comparing these programs* revealed that the information available about their effectiveness was shockingly inadequate. Only very rough judgements about their relative effectiveness could be reached. Information on costs per participant was often unavailable and had to be estimated. Data on the characteristics of participants (even, in some cases, their age and sex) were inadequate. Information on what happened to the participants during and after the Program--gains in educational attainment, employment experience, earnings history, and so forth--was extremely limited. Estimates had to be pieced together from a variety of sources."

Source: U. S. Department of Health, Education, and Welfare, Office of the Assistant Secretary for Program Coordination (1966-1970), Program Analysis: Selected Human Investment Programs, October 1966, p. 3.

*Vocational Rehabilitation
Adult Basic Education
Work-Experience and Training
Vocational Education
Elementary and Secondary Assistance
for Educationally Deprived Children

Krulee, Hetzner, and McHenry in discussing the Follow Through Program commented:

"There was dissatisfaction with the decentralized method of field supervision that was being employed, using local and often part-time supervisors. With this arrangement, it was hard to maintain quality control over training and data collection or to provide for uniformity in data collection procedures among projects."

Source: Gilbert K. Krulee, William A. Hetzner, and Emmet J. McHenry, Program Evaluation: The Follow Through Experience, p. 30.

Random Selection of Experiment Participants

GENERAL STATEMENT OF THE PROBLEM

Ideally, the design of a social experiment, like that of many medical experiments, should provide for randomly selected individuals, some of whom would participate in the experiment and others who would serve as a "control" group. Rarely has it been possible to achieve this condition in the past because citizens and politicians normally want to make a new, apparently useful service available to all who need or want it. Some experienced evaluators have suggested that control groups not be used because of the procedural problems associated with using them. However, if such control groups are to be used, random selection should be sought. Otherwise, the validity of any findings of an evaluation are greatly reduced.

ILLUSTRATIVE PAST EXAMPLE

One of the criticisms made concerning Head Start was that the research was not truly experimental, in that neither treatment nor the control groups were randomly drawn. In addition, it was found in the El Salvador ITV experiment that political factors mitigated against randomly assigning students to the innovative program.

ANTICIPATED ESCD ANALOGUE

Since volunteer sites and people are being recruited to participate in many of the ESCD programs, the evaluation of the ESCD program must include a procedure that will minimize the adverse effects resulting from the lack of random selection.

APPENDIX #7

Random Selection of Experiment Participants

"Ideally, the researcher would also like to assign students, teachers and instructional conditions randomly throughout the school system. But a school system is not a laboratory, and experimental control was not a high priority for Salvadoran administrators who were trying to deal with a major educational reform. While the researchers would have preferred to vary one element of the Reform at a time, in order to isolate the effect of ITV from other Reform elements, the political necessity of introducing the Reform as a package outweighed that preference."

Source: Robert C. Hornik, et al., Television and Educational Reform in El Salvador, Final Report, Institute for Communication Research, Stanford University, August, 1973, p. 274.

"It would require numerous lengthy chapters to describe adequately the methodological shortcomings in the Head Start (and Follow Through) Planned Variation research designs. . . The basic design of the Planned Variation studies is not a classical 'experimental' design, since sites throughout the country were not randomly assigned to preschool models and children were not randomly assigned to model and 'control' classes within sites. This problem is what Datta (1971) has called 'the soft underbelly of the mega-studies; the inability to over-recruit and assign children to treatments or sites to treatments at random or in keeping with sound design principles' (p.5). Without random assignment it cannot be assumed that groups in different models are comparable upon entry into the program and, therefore, that apparent model effects are attributable to true effects of the models rather than to characteristics of self-selected participants."

Source: Joan S. Bissel, Planned Variation in Head Start - Follow Through.

Unrepresentative and Inadequate Sampling

GENERAL STATEMENT OF THE PROBLEM

Because data collection is both one of the most costly and time consuming tasks associated with program evaluations, such evaluations usually employ statistical samples covering only part of the total target population. However, if these samples are either too small or in some way unrepresentative of the total experimental group, conclusions drawn from these sample data cannot be accepted as reliable indicators of a program's relative effectiveness. To avoid these problems, an adequate sampling procedure should be an integral component of the contractor's total evaluation design. As a minimum, sampling should be simple, capable of easy replication, and should be stratified.

ILLUSTRATIVE PAST EXAMPLE

The Head Start program was frequently criticized relative to its sampling procedure, in part because stratified random sampling was not employed, which would have recognized differing program center characteristics.

ANTICIPATED ESCD ANALOGUE

Available information describing the manner in which the three ESCD Demonstration Components selected their sites and participating citizens is sketchy. It is not possible to tell whether the participants will represent a reasonable cross-section of the total population of each of the three regions. Even more important, the Phase I proposals of the three Phase II contractors are vague in saying whether they will conduct all of their observations on a complete enumeration basis covering all project participants or whether they will use sampling, and if they do intend to use sampling, what they will do to ensure that their samples are adequate and representative.

Unrepresentative and Inadequate Sampling

"In our judgement the Westinghouse-Ohio procedure for selecting the sample seriously limits the relevance of the study to overall policy. This judgement rests on two arguments.

First, stratified random sampling, not simple random sampling, should have been used. Contrary to what the authors claim, there were reasons to believe that centers' characteristics would be related to their effectiveness. Many studies show that region and locality strongly relate to educational achievement--the Coleman Report is a classic example. Schools' ethnic and social class composition may also seriously affect an educational program. Furthermore, and again contrary to what the authors claim, the center characteristics were not entirely unknown. . . .

Stratified random sampling would have been entirely consistent with the RFP. . . .

Many kinds of centers were almost unrepresented in the sample. The problem is particularly acute in the analyses of full-year centers. . . ."

Source: Marshall S. Smith and Joan S. Bissell, "Report Analysis: The Impact of Head Start," Harvard Educational Review, Vol. 40, No. 1 (February 1970), p. 65.

Open-Ended Interview Techniques

GENERAL STATEMENT OF THE PROBLEM

Frequently much of the information needed for evaluations must be obtained from open-ended interviews before, during, and after the demonstrations have been undertaken. Planning, conducting, and analyzing the results of such interviews require considerable skill, especially when the respondents are of much different cultural backgrounds than the interviewers.

ILLUSTRATIVE PAST EXAMPLE

Critics have commented adversely on the reliability of both the Head Start and Follow Through evaluations. In the case of the Head Start evaluation, interviews which employed open-ended questions, often elicited highly subjective and possibly inaccurate information.

ANTICIPATED ESCD ANALOGUE

The ESCD performance evaluations will be importantly dependent upon the personal impressions of relatively unsophisticated program participants covering subtle, subjective matters. For this reason, well designed and conducted interviews--perhaps supplemented by written questionnaires--will be needed to obtain as accurate a record as possible of the reactions of program participants.

Open-Ended Interview Techniques

"The (Headstart) interviews consisted largely of open-ended questions which were employed by the field staff in pursuing the causes of changes that had been identified. Some of the questions in the interview schedules elicited information based on the respondent's personal judgement, which could be classed as highly subjective. Also, much of the information elicited was based on the recall of respondents who may have inadvertently provided inaccurate data."

Source: Kirschner Associates Incorporated, A National Survey of the Impacts of Head Start Centers on Community Institutions, (prepared for HEW), 1970, p. 24.

". . . obviously, one ought to be interviewing parents as well as teachers; but what did one want to learn from these interviews? Everyone could agree that the attitudes of parents and teachers had some influence on children and on classroom outcomes but it was much more difficult to agree on a model or theory about these influence processes or on a statement of hypotheses that were to be tested with the aid of these data. Similarly, it seemed obvious that one wanted to collect data relating to institutional change; again, it was conceptually very unclear what data should be collected and what hypotheses were to be tested with the aid of these data. In addition to these problems of conceptual clarification, there were the related problems of developing, pretesting, and refining a suitable set of instruments for the collection of data."

Source: Gilbert K. Krulee, et al., An Organizational Analysis of Project Follow Through (prepared by Northwestern University for the Department of Health, Education, and Welfare), 1973, pp. 207-208.

LESSONS LEARNED AND PROBLEMS ANTICIPATED SHEET #10

Comparability of Experimental and Control Groups

GENERAL STATEMENT OF THE PROBLEM

The validity of the evaluation of the effectiveness of a new social service program depends critically upon the comparability in the socio-economic characteristics of the experimental groups and the characteristics of the control group. Very frequently in the past, evaluation studies have been severely attacked on the grounds that the two groups were sufficiently different in their characteristics to render invalid any conclusions reached in the evaluation studies.

ILLUSTRATIVE PAST EXAMPLE

The evaluation studies of the Job Corps Program and Upward Bound have been criticized on the grounds that the experimental and control groups differed. Different evaluations of the former program reached vastly different conclusions relative to its effectiveness, due principally to the use of data drawn from different control groups. In the latter case, the racial composition and attrition rates of the experimental and control groups were allegedly unrepresentative.

ANTICIPATED ESCD ANALOGUE

Neither the published plans of the three Demonstration Sites nor the Phase I proposals of the three bidders selected for the Phase II design competition have much meaningful to say about control groups. For instance, available information does not make it clear whether the ESCD program will allow the selective exclusion of individuals within treatment communities. The probability that the before-the-experiment baseline data on the groups that will participate in the demonstrations is likely to be incomplete and of a low quality make this relative inattention to control groups an even more serious hazard. Clearly, how well the three bidders address this problem in their Phase III design plans will be a critical consideration.

APPENDIX #10

Comparability of Experimental and Control Groups

"The primary objective of the study was to compare the differential effects of the Upward Bound experience on blacks and whites in educational attitudes, academic motivation, and academic performance. Although the students came from a national representative sample of 21 Upward Bound centers, they cannot be considered to be representative of all Upward Bound students (1) because only black and white students were included, and (2) because of attrition of subjects over the study period. Two less-than-adequate types of control groups were used in the analysis."

Source: McDill, E.L., McDill, M.S., and Sprehe, J.T., Strategies for Success in Contemporary Education, John Hopkins Press, Baltimore, Md., 1969.

"An analysis by an economist outside of the agency (OEO) estimated the cost-benefit ratio to be about 1.5:1: Job Corps Analysts came up with 4:1: The General Accounting Office found that their calculations yielded 0.3:1. Subsequent analysis of the three estimates indicated that the differences stemmed largely from using different sources to establish the effectiveness of the program. In short, the benefits in increased income estimated for Job Corps trainees varied widely depending upon which group of young people were used as 'control groups' for estimating differences."

Source: Robert Levine, "Evaluating the War on Poverty," in On Fighting Poverty (J.L. Sundquist, Editor), Basic Books, New York, 1969.

Knowledge Transfers Between Experimental and Control Groups

GENERAL STATEMENT OF THE PROBLEM

Sometimes the evaluation of theoretically well designed experiments can be greatly complicated in practice if target groups and control groups are co-located with the result that the former unintentionally transfer new knowledge to the latter.

ILLUSTRATIVE PAST EXAMPLE

The problem of knowledge transfer was encountered in the Follow Through evaluation in which it was observed that ideas had been transferred from participants to non-participants.

ANTICIPATED ESCD ANALOGUE

It would appear that coping with these knowledge transfers among target groups and control groups would be particularly difficult for the ESCD evaluation since the participating communities are small and culturally tightly knit. In Alaska communities with total populations of as low as 150 persons are qualified to participate. In the Rocky Mountain States some participating communities have as few as 31 people and the largest has slightly less than 8,000. Under these conditions if control groups are used, very well planned and highly disciplined administered evaluations will be necessary to minimize knowledge transfers among family members, relatives, and peers.

APPENDIX #11

Knowledge Transfers Between Experimental and Control Groups

". . . SRI staff members who have been visiting Follow Through classrooms and non Follow Through classrooms have noticed that Follow Through ideas are being picked up in other classrooms in the same community. Possibility of this sort of diffusion would make it unwise to use non-Follow Through children in the same community as controls for Follow Through children: if both experimental and comparison groups make progress, the evaluation might give too pessimistic a view of the results achieved."

Source: Joseph Wholely et al., Federal Evaluation Policy, Urban Institute, 1970, p. 92.

Uncontrolled Educational System Characteristics

GENERAL STATEMENT OF THE PROBLEM

Frequently it has been difficult to compare the success of a given program applied in different geographical locations because important characteristics of the educational systems, not subject to the control of the demonstration program, were highly variable from place to place.

ILLUSTRATIVE PAST EXAMPLES

The evaluations of innovative educational programs, such as Title I and the El Salvador ITV experiment, have been criticized relative to their inability to incorporate into their design all major educational system characteristics affecting the performance of the system, such as the factor of teacher ability in Albuquerque, New Mexico Title I evaluation.

ANTICIPATED ESCD ANALOGUE

Within each of the ESCD components, the characteristics of the participating educational systems beyond the control of the ESCD program managers differ greatly. In Appalachia, comparisons in the ESCD teacher improvements will be made extremely difficult since Fredonia, New York, Erie, Pennsylvania, Coalfield, Tennessee, and Rainsville, Alabama vary greatly in such considerations as teacher qualifications requirements, salary levels, prior exposure to educational innovations, and the caliber of the local colleges participating in the ESCD program.

Uncontrolled Educational System Characteristics

". . . (in the case of Title I educational programs) individual teachers account for more of the variance in children's performance than do any other variables."

Source: Paul Rost et al., A Model for Evaluating Title I Programs, Albuquerque, New Mexico, November 16-17, 1972, p. 27.

"The results were inconclusive. Both the experimental and the control classes gained considerably during the year, but the gains tended to be larger, although not sufficiently so as to reach an acceptable level of statistical significance, in the television classes. Further examination indicated that the difference in teachers in the experimental and control sections had something to do with this."

Source: Wilbur Schramm, Instructional Television in the Educational Reform of El Salvador, Academy for Educational Development, Washington, D. C., 1973, p. 45.

"In El Salvador, as in most other places where instructional television has been used in a sophisticated way, television is simply one element in a system-wide change, almost inextricably interwoven with other changes. Therefore, when one pretends to measure the effects of the television, it is necessary to be extremely cautious about distinguishing between its effects and the effects of other elements in the system."

Source: Institute for Communication Research, Stanford University and Academy for Educational Development and U. S. Agency for International Development, Research Report No. 1 of the Study of Instructional Television in El Salvador, Design of the Study, 1968, p. 4.

Allowing for Different Quality Levels of TV Programming

GENERAL STATEMENT OF THE PROBLEM

Past evaluations of educational TV programming have demonstrated that the quality level of TV programming can greatly influence efforts to gain and hold television audience interest. This is true whatever the communications medium employed (conventional over-the-air, cable, satellite, etc.) and whatever the end-use to which the television is put, e.g., entertainment, education, other public service, etc.

ILLUSTRATIVE PAST EXAMPLE

This problem has been encountered in both the El Salvador ITV experiment and in the Sesame Street program. It was found that uneven quality of television programming was one of the greatest problems in the El Salvador ITV experiment.

ANTICIPATED ESCD ANALOGUE

Many characteristics of the ESCD experiments, including their relative short duration, will for most applications preclude major variations in the nature of TV programming material. However, to make it possible to reach any fundamental conclusions as to the effectiveness of satellite television as an educational medium, the ESCD evaluation contractor will have to provide in his design some mechanism for estimating satellite effectiveness if different types of programming than that actually used in the ESCD programs were employed.

APPENDIX #13

Allowing for Different Quality Levels of TV Programming

"One of the great worries of the leaders of the El Salvador project, and one of the great complaints of the teachers, has been the uneven quality of the television programs. The lesson that seems to emerge from this experience is that if a country wants to "do it itself" (which has advantages over the long term), it must allow sufficient time before the project goes on the air to train production teams, let them gain experience, and test and remake as many programs as possible. For this kind of procedure, of course, more nearly adequate studio facilities must be provided than the single studio El Salvador was able to use through the first three years of the program."

Source: Wilbur Schramm, Instructional Television in the Educational Reform of El Salvador, Academy for Educational Development, Washington, D.C., 1973, p. 78.

"Sesame Street's success in terms of public popularity has made it, and the various characters who appear in the series, household words. To the television world it has demonstrated that educational TV can be commercially successful. In some cases, Sesame Street has outdrawn competing commercial programs in popularity. . . . During the 1970-71 season, the program was reaching more than six million preschool children over 250 stations. In 1970, Sesame Street won the George Foster Peabody Award for meritorious service to broadcasting. .

.....
Compared with findings from evaluation research on other compensatory education programs, Sesame Street appears encouraging. Findings to date are consistently positive and relate directly to amount of viewing time. In addition the test results suggest that transfer of learning is an outcome of the program. In attempting to account for Sesame Street's apparent success, one turns most immediately to the method of preparation of program segments."

Source: Rossi and Williams, op. cit., pp. 167,169.

Separating Program from Non-Program Effects

GENERAL STATEMENT OF THE PROBLEM

Social service programs are not operated under the rigorously controlled conditions of a physical science laboratory. Target group populations exercise much personal freedom and continually interact with other groups in society. Historically, one of the most difficult problems in evaluating the effectiveness of any social service program has been to separate out the impacts of the program, per se, from the impacts of other public and private forces concurrently impacting on the target group.

ILLUSTRATIVE PAST EXAMPLE

Both education and poverty relief programs in the 1960s were often subjected to the criticism that even when improvements occurred in the socio-economic status of target groups these changes could plausibly be explained by factors other than the specific programs directed at improving these conditions. A similar condition was found to exist in the El Salvador ITV experiment. For instance, in a labor training program considerations such as increasing maturity might often lead to a higher wage scale of the participating individual.

ANTICIPATED ESCD ANALOGUE

The target groups at the three ESCD test sites will be exposed to the satellite educational material for only a fraction of their total waking hours (11% weekly). For instance, junior high students, the largest target group in the Rocky Mountain states, will be exposed to less than 3 hours weekly of satellite programs. The time exposure of target groups to the family, to peer groups, and to other community forces will be many times larger than the time exposure to the ESCD satellite programming.

Separating Program from Non-Program Effects

"If ITV had been the only aspect of Third Cycle Reform, these data would be an indication that ITV in El Salvador was cost effective. However, other inputs besides ITV clearly did contribute to the Reform of the Third Cycle. Third Cycle teachers were helped to improve their skills through retraining. Attempts to improve the supervisory system, new recreational facilities, and technical assistance also were part of the Third Cycle Reform. Unfortunately, however, it is impossible to assess exactly how much of the costs of these Reform inputs should be attributed to the Third Cycle. Therefore it cannot be said with certainty whether the ITV Reform for the Third Cycle was in fact cost effective.

Source: Wilbur Schramm, Instructional Television in the Educational Reform of El Salvador, Academy for Educational Development, Washington, D. C., 1973, p. 53.

". . . so much change is occurring spontaneously or in response to larger movements in the society as a whole that it is difficult to separate out the changes which are the result of a particular program from those which have occurred in response to other events."

Source: Peter Rossi and Walter Williams, Evaluating Social Programs: Theory, Practice, and Politics, Seminar Press, New York, 1972, p. 29.

"How much of the change in the wage income of a trainee should properly be attributed to his training? . . . Young workers just entering the labor force typically experience considerable unemployment and only low wages, partly as a result of laws that prohibit them from taking certain jobs. More important, perhaps, is the fact that a youth is trying out jobs in search for work that appeals to him, a process that often leads to unemployment. As he ages, his wages and employment increase. If a training program has a larger number of youths, much of the observed increase of income of the trainees can be attributed to this maturation process."

Source: Thomas K. Glennan, Jr., Evaluating Federal Manpower Programs: Notes and Observations, RAND Corporation, RM-5743, 1969.

Detecting Delayed Intended Impacts

GENERAL STATEMENT OF THE PROBLEM

Frequently many of a program's most important impacts are unlikely to surface until several or more years after both the demonstration program and the evaluation of that program have been completed. Historically, it has been very difficult for evaluation task forces to detect such delayed impacts with any reliability.

ILLUSTRATIVE PAST EXAMPLE

Evaluations of the HEW Title I educational programs often had difficulty in devising methods to evaluate what the long-term future earnings of a student would be through the use of tests and other measurement devices.

ANTICIPATED ESCD ANALOGUE

Any impacts--even direct first-level impacts-- of many segments of the ESCD experiments are unlikely to become apparent, at best, until three or four years after the experiments have been completed. For instance, the Federation of Rocky Mountain States will be using the ESCD program in an effort to further the career goals of junior high school students. However, these students will not be making any meaningful career decisions until they have completed high school, which will be approximately four years after the conclusion of the ESCD experiment.

Detecting Delayed Intended Impacts

"It would take at least a decade to obtain even preliminary indications of the impact of Title I on the earning capacity of children now in elementary school. We are therefore forced to use interim measures of the educational progress of deprived children, such as test score gains, which may be presumed to be correlated with future earning capacity."

Source: U. S. Department of Health, Education, and Welfare, Office of the Assistant Secretary for Program Coordination (1966-10), Program Analysis: Selected Human Investment Programs, October 1966, p. 32.

Anticipating Unplanned Side Effects

GENERAL STATEMENT OF THE PROBLEM

The unplanned side effects of major social service programs are often more momentous than are the intended impacts of the programs. For this reason new interdisciplinary fields of study--with various titles (technology assessment, social impact analysis, etc.) and using a new collection of methods and techniques--have grown up with the express purpose of predicting unplanned side effects of major complex programs.

ILLUSTRATIVE PAST EXAMPLE

Major social programs, such as the FHA insurance program and communication satellite programs, have had notable side effects. The impact of the FHA program on segregation, urban sprawl, and urban migration, for instance, were completely unanticipated.

ANTICIPATED ESCD ANALOGUE

Because the ESCD program brings together some of today's most modern technologies with some of the nation's most isolated cultures, it would be most surprising if significant unanticipated side effects, both favorable and unfavorable, did not occur. The ESCD RFP takes explicit account of this possibility when it states: "Some parts of this work may be at a level of uncertainty and difficulty which require state-of-the-art capabilities in technological impact assessment. This technology assessment capability would be directed toward identifying, describing, and measuring the total impact consequences of the ESCD experiment as opposed to measuring only its relative success in achieving its professed mission objectives. These total consequences would embrace indirect as well as direct effects, delayed as well as immediate effects, . . ." The winning Phase III evaluator contractor's design proposal should describe in some detail how he will satisfy this requirement.

Anticipating Unplanned Side Effects

"In the past, side effects have received little attention from program managers or policy makers. The effects of FHA insurance policies and practices on segregation, urban sprawl and the decline of the central cities, for example, went virtually unnoticed for years. Similarly, it came as somewhat of a surprise to many of those who managed public housing that their projects often seemed to magnify problems of delinquency and family disintegration. Only recently have the socially destructive aspects of public assistance programs been acknowledged. It is the task of managers and evaluators to assemble fair assessments of both positive and negative side effects."

Source: Joseph S. Wholely, et al., Federal Evaluation Policy, Urban Institute, 1970, pp. 34-5.

"The use in particular of communication satellites by less developed countries presents several interesting considerations. Basic to any discussion is a recognition first of the difficulties characterizing interactions of technology transfer with respect to cultural relationships. . . . One writer notes, 'To a very large extent, the serious unemployment and underemployment that characterizes the less-developed nations of South America today is due to an indiscriminate application of production techniques which have been developed in and for the economic, social and cultural context of highly industrialized nations, where skill levels are high and labor tends to be scarce.' This object lesson substantiates the need for very careful planning for any sort of technology transfer from advanced to less developed nations. The very gap which separates the two implies a gap in planning frameworks. Hence the application of new communications afforded by satellites may also be filled with relatively unpredictable variables or effects."

Source: Bert Cowlan, et al., Broadcast Satellites for Educational Development: The Experiments in Brazil, India, and the United States, Academy for Educational Development, 1973, pp. v-vi. (Quotation: David L. Morse, "Unemployment: Bitter Burden of Millions in South America," Focus, Feb., 1973, p. 4.)

Integrating Incommensurables

GENERAL STATEMENT OF THE PROBLEM

Program managers, politicians, and action-oriented people in general, at the conclusion of a demonstration program like a simple, straight-forward answer to the questions: "Did the program work, did it accomplish what it set out to do, should we do more or less of this type of thing in the future?" In practice, it is difficult methodologically to provide an unequivocal answer to this type of question. One reason is that a major program often is intended to accomplish different things. Even if quantitative measures can be developed to estimate benefits and costs separately in various target areas, it is hard to integrate these separate answers into a composite performance measure for the program as a whole.

ILLUSTRATIVE PAST EXAMPLE

The evaluators of the El Salvador ITV experiment found it extremely difficult to compare through conventional means, such as cost-benefit analysis, the benefits of the ITV educational experiment as compared to, for example, the value of different crops.

ANTICIPATED ESCD ANALOGUE

The ESCD program is being conducted to answer the fundamental question: Are satellite telecommunications a more or a less effective means of spending public dollars for improving the educational opportunities and achievements of isolated, rural people as compared to alternative technologies and programs? However, each of the three ESCD components has been designed to satisfy particular local needs. Even within a given geographical component, there is much variation in objectives. For instance, it will be conceptually difficult to derive a composite measure of the achievement of the Federated Rocky Mountain States program since junior high school students, interested professional people, parents, teachers, and members of the community are all participating as target groups in the ESCD experiments.

APPENDIX #17

Integrating Incommensurables

"Cost-benefit analysis has proved to be a powerful tool in raising the efficiency of industries throughout the world. Indeed unless one weighs the measured benefits derived from any endeavor against the costs of that endeavor, he can never be certain that the funds invested in the endeavor might not have been better directed in some other enterprise.

Along these lines one might ask, have the benefits of the Educational Reform in El Salvador been worth the costs? Would the funds invested in Reform ITV have been better invested in say, agricultural development? First of all there is the difficulty of putting a price tag on human learning, or even quantifying it satisfactorily. Without any unit value on education it is impossible to compare educational output with that of cotton or coffee beans, on which a monetary value can easily be placed. The only readily available standard which then can be used in making even an extremely crude cost-benefit analysis is that of the pre-existing Salvadoran educational system. That is, how do the costs and benefits of Reform ITV compare with those of the Old System?

Source: Wilbur Schramm, Instructional Television in the Educational Reform of El Salvador, Academy for Educational Development, Washington, D.C., 1973, p. 52.

LESSONS LEARNED AND PROBLEMS ANTICIPATED SHEET #18

Identifying Possible "Hawthorne" Effects

GENERAL STATEMENT OF THE PROBLEM

Programs of relatively short duration are subject to what has been called the Hawthorne effect. This means simply that participants in an experiment behave differently merely because they know that they are involved in an experiment. Detecting the extent to which this effect is operative, how significant it is, and separating its effect from the effects of the program, per se, is often a difficult analytical challenge in any evaluation exercise.

ILLUSTRATIVE PAST EXAMPLE

Evaluations of the New Jersey Income Maintenance Program have been criticized on the basis that they did not adequately cope with the fact that the program's short duration would cause participating families to react in atypical ways to the receipt of the income supplements.

ANTICIPATED ESCD ANALOGUE

For several reasons the ESCD demonstrations are likely to be subject to Hawthorne-type effects. The target populations, having lived in isolated, rural environments all of their lives, will probably find the exposure to satellite educational TV a somewhat strange, fascinating experience. Some of them have had no prior exposure to any kind of TV. These initial reactions are likely to endure for a good part of the total experimental period where, as in some cases in Alaska, the total number of exposures will be only 16 or 17. The challenge for the Phase III evaluation contractor will not only be to detect the presence and magnitude of Hawthorne effects, but to ascertain how soon and how fast they wear off.

Identifying Possible "Hawthorne" Effects

David Kershaw, Field Director for Mathematica, the organization that conducted the evaluation of the New Jersey Negative Income Tax experimental program, expressed this problem as follows:

"One of the most common criticisms of the New Jersey experiment is that it will not last long enough to allow families to react to the receipt of the transfers in a 'natural way,' that is, in the same way they would were the payments part of a permanent national income maintenance program. Within the larger area of uncertainty surrounding an experiment of relatively short duration are a number of related, specific criticisms: (1) the families will feel like guinea pigs and will react simply because they are part of an experiment ('Hawthorne effect'); (2) they will react to the aims of the experiment and use the money in atypical ways ('funny money'); (3) they will develop a conditioned pattern of response to the quarterly interviews during the experiment and thus their answers will be a function of the fact that they have been interviewed previously; (4) the relatively short duration of the experiment will make responses unnatural and atypical (they will be either getting used to it or looking toward the termination date)."

Source: David N. Kershaw, "Issues in Income Maintenance Experimentation" in Rossi and Williams, Evaluating Social Programs: Theory, Practice, and Politics, p. 237.

Duration of Initial Effects

GENERAL STATEMENT OF THE PROBLEM

It was noted previously (#15) that certain intended effects of a program will not surface until several years after the program, and its evaluation, have been concluded. The converse of this point is that sometimes, apart from "Hawthorne effects," a program's initial effects will disappear within several years time. The methodological challenge to those conducting evaluations is to find a reliable method that will predict in a given case the extent to which this "reversion" effect will operate.

ILLUSTRATIVE PAST EXAMPLE

Reversion effects presumably occurred in the case of Head Start, in which it was found that over a long period of time the cognitive and affective differences between participating and non-participating students diminished.

ANTICIPATED ESCD ANALOGUE

The only reliably method of detecting lost-learning or reversion effects is by systematic follow-up studies covering the status of participants in demonstration studies several years after the demonstration has ended. Because the evaluation contractor will have only a relatively short period in which to gather data following the conclusion of the ESCD demonstration programs, this contractor will have to employ a highly responsive methodology to detect reversion effects.

Duration of Initial Effects

"The general finding is that Head Start participants record gains on various cognitive and affective measures when tested at the beginning and end of Head Start. However, the relatively few follow-up studies performed generally show that, by the middle of the first year of regular school, any differences which may have been observed between the Head Start and control groups immediately after Head Start have disappeared."

Source: Walter Williams, Social Policy Research and Analysis: The Experience in the Federal Social Agencies, American Elsevier Publishing Co., New York, N. Y., 1971, p. 106.

Applying Discount Rates to Projected Long Term Benefits

GENERAL STATEMENT OF THE PROBLEM

To translate projected long term benefits of social service programs, especially those involving young people, into quantitative, dollar estimates, it is necessary to convert future benefits into current values by applying discount rates. Although frequently practiced, this procedure has generated a continuing professional controversy relative to its validity.

ILLUSTRATIVE PAST EXAMPLE

The evaluation of the OEO Job Corps Program was criticized on the use of discount rates in that by merely varying these rates the estimated benefits of the program could be drastically altered.

ANTICIPATED ESCD ANALOGUE

Since most of the participants in ESCD programs will be persons who are either now or will be entering the labor force, some procedure will be needed to translate various projected economic benefits into the form of a common denominator. Discounting likely will be the technique used. The ESCD evaluation contractor will have to practice a high degree of professional expertise to avoid some of the pitfalls associated with the more simplistic approaches to discounting employed in the past.

Applying Discount Rates to Projected Long Term Benefits

"The extrapolation of program effects measured shortly after termination from the program into long-term benefits is largely based on unfounded assumptions. There is little theoretical basis and even less empirical evidence to justify the extrapolation of the influence of a four-to-six month program, 10 to 20 years into the future--not to mention 30 to 40 years. Further, we have demonstrated the significant influence of discounting and yet there is considerable disagreement as to whether the appropriate rate is 3 percent or 9 percent--a factor of three to one. . . . Doubling the discount rate (on Job Corps programs) from 5 to 10 percent reduces the present value of the benefits by more than 60 percent."

Source: Resource Management Corporation, Evaluations of the War on Poverty, the Feasibility of Benefit-Cost Analysis for Manpower Programs, (prepared for the General Accounting Office), March 11, 1969, pp. 120,129.

LESSONS LEARNED AND PROBLEMS ANTICIPATED SHEET #21

Resistance to Evaluation

GENERAL STATEMENT OF THE PROBLEM

No one likes to be evaluated publicly. For this reason Program Directors and participating communities have often used various strategies to vitiate evaluation efforts, e.g., keeping even measurable program goals fuzzy, deliberately limiting the flow of readily available information to the evaluation task force, etc. These politically induced impediments--especially when coupled with the inherent conceptual difficulties of defining certain program goals and the genuine lack of certain types of data--have severely hampered many past efforts at evaluation.

ILLUSTRATIVE PAST EXAMPLE

The evaluation of the Title I program and Head Start often encountered the problem of administrators who were cautious of the program, who lacked interest in it, and who feared that they would be damaged through an evaluation.

ANTICIPATED ESCD ANALOGUE

Only time will tell how much this type of problem will occur with the ESCD experiments. Knowing human nature, and in view of the historical record on other programs, it would be surprising if the ESCD Phase III evaluation contractor were spared this problem. The winning bidder's proposal should show awareness of this potential hazard, and produce a credible strategy for eliciting the cooperation, as opposed to a resistance, of the program participants.

Resistance to Evaluation

"Because states lacked interest and incentive--and, in some cases, because they did not want to circulate (Title I) information that they conceived to be damaging to themselves in some way--they did not send in the required data."

Source: Joseph Wholely et al., Federal Evaluation Policy, Urban Institute, 1970, p. 34.

" . . . the agencies and the districts we work with were somewhat shy about evaluations and I'll tell you why they were shy. They felt that evaluations or testing have traditionally been used to be critical of or to sabotage, a program. There is still some feeling that 'You're not measuring me to determine how well I'm doing, but you're measuring me in order to criticize.'"

Source: Introductory Remarks by Dr. Wilson E. Riles, Director of the Division of Compensatory Education, California State Department of Education, in introducing Jeaneda H. Nolan's talk entitled "A Report on the Evaluation of the State Preschool Program Contrasted with the Westinghouse Report on Head Start," presented to the California State Board of Education, June 1969.

Evaluator-Target Group Interface

GENERAL STATEMENT OF THE PROBLEM

An unrecognized communications gap between the sophisticated evaluation researcher and the unsophisticated target group sometimes leads the former to reach erroneous conclusions relative to the achievement levels of the latter.

ILLUSTRATIVE PAST EXAMPLE

In the evaluation of Sesame Street, it was often difficult for an evaluator to test a student's ability on a level which the child could fully comprehend.

ANTICIPATED ESCD ANALOGUE

There will be few, if any, very young children participating in the ESCD experiments. However, there still could be an unrecognized, culturally induced communications gap, and its untoward consequences, arising from semantic misunderstandings. These could develop from the fact that the people participating in the ESCD demonstration are from very isolated, rural regions, whereas most of the staff of the evaluation contractor will have urbane, metropolitan backgrounds.

Evaluator-Target Group Interface

" . . . among the arsenal of tests available for use with young children, we found the situation appalling. Take, for example, items such as: 'Put the green marbles in the square box.' If the child gets it wrong, is it because he does not know the meaning of 'green,' 'marbles,' 'square,' or 'box?' Or, what if the child is given a stimulus triangle and asked if she can find another one just like it embedded in a larger drawing? The child is asked to trace around it. If she cannot perform the task correctly, does this mean that she did not find it, that she has poor motor coordination, or that she does not want to play games with the tester? In one well-known standardized test for five-year-olds, we found that the percentage passing national norms for a particular counting item were lower than we had obtained at pretest with four-year-old disadvantaged children. Then we looked at the administration manual, where we found the instruction for the children was: 'We are now going to play a number game. Look at the pictures of the fish bowls at the top of the page. Let's pretend that there are five tables in your class and each table has a fish bowl on it. See the large letters above the fish bowls. These letters are A, B, C, D, and E. Listen carefully. In which bowl do you see just three fish?' What happens to a child who does not understand words like 'pretend' and who has never seen a fish bowl?"

Source: Samuel Ball and G. A. Bogatz, Research on "Sesame Street:" Some Implications for Compensatory Education, p. 14.

Midstream Changes in a Program's Characteristics

GENERAL STATEMENT OF THE PROBLEM

Sponsors of social service programs, and those who manage them, are primarily interested in servicing the specific target groups the programs have been designed to help, rather than in developing an elegant data base for a formal evaluation of the programs. Therefore, directors of new programs want the option to modify midstream the objectives of the programs, and even more the nature of the program's activities. Such modifications usually improve the effectiveness of a program over time. However, these modifications also figuratively put the evaluation task force in the position of shooting at a moving target and of making the evaluation more difficult.

ILLUSTRATIVE PAST EXAMPLE

The history of programs such as urban renewal in the United States and the ITV experiment in El Salvador illustrate the difficult task of evaluating midstream changes in a program's characteristics. In the latter case, new curriculum and materials, and retrained teachers made it difficult to evaluate the effectiveness of the original program.

ANTICIPATED ESCD ANALOGUE

It has been our understanding that frequent changes in objectives and planned program activities have characterized the three ESCD components to date. The design proposal of the winning Phase III contractor should have a built-in response mechanism to react to further program changes after the demonstration exercise and the formal evaluation of it begin.

Midstream Changes in a Program's Characteristics

"The old Housing and Home Finance Agency, before and after its transformation into a Cabinet Department in 1966, constantly experienced major changes. While the omnibus housing acts of 1949, 1954, 1956, 1961, 1964 and 1968 are most often remembered and cited, the fact is that either the organization, responsibilities, policies or programs were altered in virtually every year since the federal government first made an in-depth commitment to play a role in housing and urban problems . . .

.
the department has not carried out objective and useful evaluation to any considerable extent. A prime reason is that legislation does not require that urban renewal be evaluated."

Source: Joseph Wholely, et al., Federal Evaluation Policy, Urban Institute, 1970, pp. 30-1.

"Comparisons such as the above were more difficult in later years because, beginning with the second year, the new curriculum and new materials were introduced into all classes, and a number of classes, without television were taught by teachers who had been retrained. Therefore, a comparison of the groups could show clearly neither the differences attributable to change of system nor to use of television."

Source: Wilbur Schramm, Instructional Television in the Reform of El Salvador, Academy for Educational Development, Washington, D.C., 1973, p. 42.

Local Diversity in Demonstration Programs

GENERAL STATEMENT OF THE PROBLEM

It is often impossible to arrive at a valid national-level, consensus type evaluation of a program that allows for great diversity at the local level in objectives, activities, and administrative procedures.

ILLUSTRATIVE PAST EXAMPLE

The variety in Head Start centers in matters such as curriculum modification, priorities, the independence of teachers, and willingness to change made it difficult to evaluate the Head Start program consistently on a nation-wide basis.

ANTICIPATED ESCD ANALOGUE

The rather abbreviated available description of each of the planned three ESCD components indicates not only a great diversity among the three components when each is viewed totally, but also seems to indicate that within any given program that a real effort is being made to respond to local needs as they are visualized locally. The winning ESCD Phase III evaluation contractor's Design Plan should deal with this situation in two ways. First, he should recognize the local diversity among programs and evaluate each of them on its own merits. Second, he should provide some plan for arriving at consensus-type conclusions concerning the efficacy of the total program for top-level planners and managers.

Local Diversity in Demonstration Programs

"The learner, the program, and the teacher function in an educational setting which has its own institutional arrangements. Systems vary in the degree to which they are open to change and willing to modify the curriculum, willingness to reorient resources and change priorities. Teachers in these systems vary in the degree of independence as well as skill and morale. In other words, we have in the educational system things of tremendous heterogeneity. . . . Thus compensatory educational programs vary from community to community as well as within communities. It is difficult, if not impossible, to expect uniform gains."

Source: Lois-ellin Data, "Are Head Starts Heterogeneous" Paper presented at the 1969 American Psychological Association annual meeting, San Francisco, p. 7.

". . . there were no uniform or standardized procedures adhered to between various programs to insure that the evaluation would be attempting to assess those factors which programs shared in common. For example, the various centers employed somewhat different goals, treatments, and program procedures, thus masking between and within center differences. Some centers were in operation for two hours per day whereas others were in operation for four hours; some centers were only in operation for two months whereas others were in operation for eight or nine months out of the year. . . . Despite these differences the programs were all evaluated as if they were similar; . . ."

Source: Gerald James Gladkowski, Another Look at Compensatory Education, 1971, p. 9.

Replicability of Findings

GENERAL STATEMENT OF THE PROBLEM

Evaluations of demonstration programs have often lost sight of the fact that a demonstration program is intended to provide a testing ground for a much wider application of the program in somewhat different environments and under somewhat different circumstances. Rarely has sufficient attention been paid to identifying both the common characteristics that would prevail in different environments and also those that would have important bearing on the chances of success or failure of the program when applied more widely.

ILLUSTRATIVE PAST EXAMPLE

The Westinghouse evaluation of the Head Start program was criticized for including too few learning centers to make it possible to identify replicable features of the program which would make for success or failure of similar compensatory programs if they were applied in other communities.

ANTICIPATED ESCD ANALOGUE

The ESCD experiments, involving rurally isolated people in the Rocky Mountain States, Appalachia, and Alaska, are intended principally as a pilot test to determine how similar programs might be applied in other underprivileged regions, but with different cultural, ethnic, and geographical characteristics. Little, if anything, in the Phase I proposals of the three winning bidders gave much evidence as to how they would seek to make their findings relevant to other communities that might consider using satellite TV in the future. Among other possible users might be the French-Canadians in northern Maine, Indian people in the Dakotas, Mexican-Americans in the southwest, and other socially deprived people who are somewhat less geographically isolated--for instance, certain rural residents of Mississippi and Louisiana. The three Phase III Design proposals should be examined carefully for their adequacy on this point.

Replicability of Findings

"A successful research and development effort implies replicable results. This suggests that the search for features which may influence the success of a program should center on those who are not unique to a particular center. We do not for a moment minimize the importance of an exciting 'group leader' or uniquely trained specialist. But if there are only half a dozen of these unique people in the country, and if a realistic constraint of the program is that it must be carried out at the local level at hundreds or even thousands of disparate centers, then we must identify replicable features which increase program effectiveness. Thus, enough centers ought to be included in a program evaluation to isolate and determine replicable features which are desirable. The Westinghouse study included only 29 full-year centers, and, thus, did not really deal with this problem."

Source: Richard J. Light and Paul V. Smith, "Choosing a Future: Strategies for Designing and Evaluating New Programs," Harvard Educational Review, February 1970, p. 7.

7
EXPERT SOURCES CONTACTED (OFTEN BY TELEPHONE)

IN THE PREPARATION OF THIS REPORT

Blaschke, Charles ... Educational Turnkey Systems
Boechman, Margaret ... Johns Hopkins University
Campbell, Donald ... Northwestern University
Cain, Glenn ... University of Wisconsin
Corbett, Joseph ... Department of Health, Education & Welfare
Cunningham, Robert ... Department of Health, Education, & Welfare
Dunn, Donald ... Stanford University
Feldmesser, Robert ... Educational Testing Services
Ingle, Henry ... Academy of Educational Development
Jamison, Dean ... Educational Testing Services
Jamison, Marshall ... Independent Consultant
Keeting, Daniel ... Johns Hopkins University
Luenberger, David ... Stanford University
McDill, Edward ... Johns Hopkins University
Mosley, Steven ... Academy of Educational Development
Ryan, Thomas ... Carleton University
Spiegelman, Robert ... Stanford Research Institute
Sprague, David ... State Department (AID)
Watts, Harold ... University of Wisconsin
Weiss, Carole ... Columbia University
Wholey, Joseph ... The Urban Institute
Williams, Walter ... University of Washington